CEFR 読解指標に基づく日本語例文自動分類 Web アプリケーションの精度向上に向けた取り組み

Attempt to Improve Accuracy of Web Application to Automatically Classify Japanese Example Sentences Based on CEFR Reading Comprehension

CAO HOAI GIANG[*1], 宮崎 佳典[*2],谷 誠司[*3]
CAO HOAI GIANG[*1], Yoshinori MIYAZAKI[*2], Seiji TANI[*3]
[*1] 静岡大学情報学部
[*2] Faculty of Informatics, Shizuoka University
[*2] 静岡大学学術院 情報学領域
[*2]College of Informatics, Shizuoka University
[*3] 常葉大学外国語学部
[*3]Faculty of Foreign Languages, Tokoha University
Email: nargiang@gmail.com

あらまし：近年関心を集めている Can-Do による言語能力尺度の一例に CEFR（ヨーロッパ言語共通参照枠）が挙げられ，世界の外国語教育に導入されつつある．本研究では日本語版 CEFR の読解項目を対象に，与えられた例文の分類手法を開発している．分類のための特徴量として数種類の特徴量を採用しているが，その中の文書タイプや専門性同定のために現行の版では fastText を用いており，これに対し新たに BERT アルゴリズムを適用することで予測精度の向上を試みる．
キーワード：日本語学習，例文検索，機械学習，情報検索

1. はじめに

CEFR（セファール）は外国語のコミュニケーション能力を表す指標であり，国際標準規格として欧米を中心に広く使われている(1)．現在提供されている参照枠は，英語を含めて 38 もの言語に上る．CEFR は具体的な言語能力レベルを A1 レベルから C2 レベルまでの 6 段階で設定しており，Reading，Writing，Speaking，Listening といった技能項目に対して，それぞれのレベルで言葉を使ってできることを能力記述文(Can-Do Statements，以下 CDS)で記述している．本研究では Reading の技能について研究を行っている．2017 年には CEFR を補完するものとして CEFR Companion Volume (2)が公開された．(2)では言語学習の初学習者向けに，言語能力に A1 レベルよりもさらに初級段階のレベルとして PreA1 レベルが追加された．

CDS は「文書タイプ」や「何が出来るか」などを示しているが．CDS の内容は抽象的なものが多いため，利用時には CDS そのものより具体的な例文のほうが使いやすい．例えば，A1 レベルの CDS には「身近な話題について日常の定型型の手紙やファックスを理解することができる」があるが，同 CDS に対して具体的な例文「昨年は大変お世話になりありがとうございました．本年もどうぞよろしくお願い申しあげます.」のほうが問題として利用しやすい．

日本語教育への CEFR の活用に関しては日本語の CEFR 準拠テキストコーパスが作成されていないこともあり，現時点において網羅的に研究されている例はまだ少ないと思われる．高田らは日本語の CEFR 準拠テキストコーパス作成時に必要な，例文に読解指標（CDS）を付与する労力を軽減するための日本語例文自動分類に取り組んだ(3)．現時点で CEFR の言語能力レベルは PreA1，A1，A2，B1，B2，C1，C2 の 7 段階がある．しかし，C1 と C2 レベルは母国語話者でも理解しにくい場合もあり，本研究では対象外とし，残りの 5 段階の能力で対応する 34 個の CDS を対象とする（B2 レベルの 1 個は読解能力よりも語彙力を重視しているため除外している）．

本研究では自動分類に際し，各 CDS を区別する特徴量として「専門性」，「文長」，「文書タイプ」に分けている．平川らはこれらの特徴量を全自動で抽出させる Web アプリケーションを開発した (4)．本研究ではさらに抽出精度を向上させることを目指す．

2. 先行研究

2.1 分類手法

本研究では直接の先行研究にあたる(5)の CDS 分類方法を継続して用いることとする．故その方法について先に上で述べた CDS 分類に使用する特徴量から説明していく．

文書タイプとは新聞記事，公的文書などといった文書の種類である．(6)では 7 種類を用いた．専門性とは例文が日常的であるか，専門的であるかに応じ

て値を取るものである．漢字率は JLPT の 5 レベルに定義される漢字の出現数を文書内で計算し，その出現率を計算したものである．文長とは文書の文字数，語数，文数，改行数を計算したものである．文書タイプと専門性の推定には文書から各単語の分散表現を獲得する fastText と形態素解析器である MeCab を用いて自動推定を行う．文長，漢字率は MeCab を用いて自動計算を行う．

これらを用いて CDS 分類を行うが，CDS は抽象的な内容であるものが多いことから，例文に対する CDS が一つに分類されることはあまりない．よって，CDS 分類には CDS をラベルとして，一つの例文に複数の CDS が対応するマルチラベル分類が想定された．マルチラベル分類方法には複数の 2 値分類器（SVM）を用いる1対他分類法が用いられた．分類結果の評価指標は，推測CDSが正しいことを正，否を負とし，それぞれ正の F 値，負の F 値を定義して用いる．

2.2　文書タイプと専門性の推定手法

fastText は文書から各単語の分散表現を獲得できる手法であり，文書分類にも用いることができる．(4) では facebook が提供している fastText を用いたテキスト分類ライブラリを適用し，パラメータ設定については epoch=45，lr=1. 0，wordNgrams=2 とし，他はデフォルト値を用いた．また分類時に必要な，文書の分かち書きには MeCab を使用した結果，全体の正解率は約 78. 07%であった．また，専門性の推定については2値分類を行った（専門である文書／非専門である文書）．手法については文書タイプと同じく fastText(パラメータの設定も文書タイプの推定と同じ)を用い，その分類精度として約 77. 61%が得られた．

3.　本研究の提案内容

BERT が Google の新しい自然言語処理技術で様々な分野で使用されている（(7)や(8)など）．文脈化された単語分散表現の特徴で構文解析，述語項構造解析などの分野で目覚ましい活躍を遂げており，そこで BERT を用いて文書タイプの分類を試みることにする．

3.1　日本語例文分類における BERT の利用

事前学習として東北大学における日本語事前学習済み BERT モデル bert-base-japanese-whole-word-masking (9) を使用することにした．その上でのファインチューニングとして，Mecab を用いて入力テキストの形態素解析を行い，分類ラベル情報と共に Pytorch の dataset を保持し，学習時に事前学習モデルに渡した．

3.2　学習データ

文書タイプラベル付の例文には7種類あり（記事，ニュース，公的文書，標識，通信文，使用説明，その他），合計 555 例文を今回では基礎データとして用いた．ただし，例文数が少なかったため，Google 翻訳API を用いて，自動翻訳プログラムを作成した．具体的には，他言語に翻訳し，再度日本語に戻すという方法でデータを増やし，各文書タイプの例文数を 4 倍にすることにし，合計 2775 例文を得た．

3.3　実験結果

10 交差検定で精度を確認した結果，基礎データである 555 例文の場合 77.4%であった．これに対し，増幅された 2775 例文では，結果は 79.3%に僅かに向上した．なお，Google 翻訳で増幅した例文とそれに対応する元の例文は常にセットとし，学習データそして検定データとして混用しないように配慮した．

3.4　先行研究との比較結果

(4) で用いられたデータが再現できていないため，結果を純粋には比較できないが，現時点で同等あるいは同等程度以上の精度が得られている．いずれも (4) のシステムで今回使ったデータでテストを行い，発表日当日に解析結果を詳述する予定である．

4.　まとめと今後の展望

本発表では CEFR 読解指標に基づく日本語例文自動分類 Web アプリケーションの精度向上に向けた取り組みについて紹介した．BERT により専門性，文書タイプにおいて分類の精度を上げる実験を試み，先行研究の精度と比較した．今後の展望は先行研究と同一条件で比較し，BERT の適用性を評価し，さらに専門性など他のパラメータの精度も向上させることで，日本語例文自動分類 Web アプリケーション全体の精度改善を目指す．

参考文献

(1) Council of Europe: Common European Framework of Reference for Languages: Learning, Teaching, Assessment, Cambridge University Press, 2001.
(2) Council of Europe: Common European Framework of Reference for Languages: Learning, Teaching, Assessment Companion Volume with New Descriptors, 2017.
(3) 高田 宏輝, 宮崎 佳典, 谷 誠司, 韓国人日本語学習者のための CEFR 読解指標に基づく例文分類, 韓國日本學會第 94 回國際學術大會, pp. 299-303 (2017).
(4) 平川 遼汰, 宮崎 佳典, 谷 誠司, 日本語例文自動分類による CEFR 読解指標推定支援 Web アプリケーションの開発, 情報処理学会第 80 回全国大会, pp.(4)-635-636, 2018.
(5) 宮崎 佳典, Vuong Hong Duc, 谷 誠司, 安 志英, 元 裕璟, CEFR Companion Volume に対応した日本語例文自動分類手法, 日本學報第 125 輯（The Journal of Korea Association of Japanology）, Vol.125, pp. 153-175 (2020).
(6) Huynh Nguyen Tra My, Y. Miyazaki, S. Tani, Inferring CEFR Reading Comprehension Index Based on Japanese Document Classification Method Including PreA1 Level, 教育システム情報学会 研究報告, Vol.33, No.2, pp. 63-69, (2018).
(7) 柴田 知秀, 河原 大輔, 黒橋 禎夫, BERT による日本語構文解析の精度向上，言語処理学会 第 25 回年次大会 発表論文集, pp. 205-208, (2019).
(8) 福田 治輝, 綱川 隆司, 大島 純, 大島 律子, 西田 昌史, 西村 雅史, 協調学習における評価対象テキストの自動評定, 第 18 回情報科学技術フォーラム(FIT), K-018, pp. 343-344, (2019).
(9) 東北大学における日本語事前学習済み BERT モデル https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking